# A deep multi-modal neural network for informative Twitter content classification during emergencies

Abhinav Kumar
Department of Computer Science & Engineering
National Institute of Technology Patna, India
abhinavanand05@gmail.com

Jyoti Prakash Singh
Department of Computer Science & Engineering
National Institute of Technology Patna, India
jps@nitp.ac.in

Yogesh K. Dwivedi*
Emerging Markets Research Centre (EMaRC)
School of Management, Swansea University Bay Campus
Fabian Way, Swansea, SA1 8EN
United Kingdom
Email: ykdwivedi@gmail.com

Nripendra P. Rana
School of Management, University of Bradford
Richmond Rd, Bradford, BD7 1DP
United Kingdom
Email: nrananp@gmail.com

*Corresponding Author

## Abstract

People start posting tweets containing texts, images, and videos as soon as a disaster hits an area. The analysis of these disaster-related tweet texts, images, and videos can help humanitarian response organizations in better decision-making and prioritizing their tasks. Finding the informative contents which can help in decision making out of the massive volume of Twitter content is a difficult task and require a system to filter out the informative contents. In this paper, we present a multi-modal approach to identify disaster-related informative content from the Twitter streams using text and images together. Our approach is based on long-short-term-memory (LSTM) and VGG-16 networks that show significant improvement in the performance, as evident from the validation result on seven different disaster-related datasets. The range of F1-score varied from 0.74 to 0.93 when tweet texts and images used together, whereas, in the case of only tweet text, it varies from 0.61 to 0.92. From this result, it is evident that the proposed multi-modal system is performing significantly well in identifying disaster-related informative social media contents.

**Keywords** Disaster, Twitter, LSTM, VGG-16, Social media, Tweets

## 1. Introduction

A natural disaster creates significant ecological disruption requiring extensive efforts from society to overcome and cope with them (Imran et al., 2015; Sakaki et al., 2013; Kumar and Singh, 2019). In the case of natural or man-made disasters, rescue organizations need to respond to all the affected people on time. However, this task is very challenging to the professional humanitarian communities and government agencies due to the limited information of the victims' location, massive number of calls by victims and their relatives, and prioritizing rescue operations based on the need of victims (Imran et al., 2015; Sakaki et al., 2013; Kumar and Singh, 2019; Kumar et al., 2019; John et al., 2018; Jin et al., 2015; Paul and Hariharan, 2012; Dubey et al., 2017; Shareef et al., 2018; Sinha et al., 2017; Nguyen et al., 2016). The lack of coordination among rescue organizations and supply chain actors results in significant financial and life loss (Dubey et al., 2019, 2014; Dwivedi et al., 2018; Jabbour et al., 2017). On average, 388 disasters have occurred annually from 2003 to 2012, causing an economic damage worth of 156.7 billion US dollars (Guha-Sapir et al., 2012). It is found that during an emergency, a massive amount of user-generated data is posted on social media platforms such as Twitter and Facebook. These social media platforms are used by the people to communicate at different levels, such as from person to person, person to government agencies, and government to people (Singh et al., 2017; Kumar and Singh, 2019; Nguyen et al., 2016; Kumar et al., 2017). Victims and eyewitnesses often post their status; report infrastructure damage; inform about injured people; and also ask for help through these platforms with text, images, and videos. These user-generated data produced through social networking sites are pervasive, rapid, and accessible that can be used to coordinate for helping the victims and empowering citizens to become more situationally aware at the time of disaster (Dubey, 2019; Caragea et al., 2016; Papadopoulos et al., 2017; Akter and Wamba, 2017). Several examples are evidence where social media has played a vital role in relief efforts, finding help, and potentially saving lives. For instance, in the case of Hurricane Harvey, a woman was rescued when she tweeted for help as the emergency contact number "911" was not reachable.[1] In the case of the Chennai flood in India, people asked for help by posting their message on Twitter (Singh et al., 2017).

Among the massive volume of tweets related to a disaster, some of them might be just thanking Twitter or local groups for their help. These tweets are not very useful for a humanitarian organization in their rescue work. These types of tweets are termed as non-informative tweets. The other types of tweets where people are asking for help, locating their relatives, provide information regarding infrastructure and utility damage, affected individuals, injured or dead people. These types of tweets are termed as informative tweets. It is impossible for emergency responders to manually go through each of the posts to mine informative posts to take action due to the massive volume and speed of tweets posting. This manual inspection can also take away valuable human resources from other essential tasks. Therefore, this creates an immediate need to build systems that can automatically

[1] http://time.com/4921961/hurricane-harvey-twitter-facebook-social-media/

filter the informative contents out of a large volume of social media content. The automatic classification of social media messages, especially tweet texts, is a challenging task due to their limitation in size (only 280 characters), non-standard abbreviations, and grammatical errors (Nguyen et al., 2017; Imran et al., 2015). Recent works by Caragea et al. (2016) and Nguyen et al. (2016, 2017a) explored tweet texts only to filter disaster-related informative tweets from social media. But people also post a good volume of images and videos related to disaster, which can give a lot of insight into the event. A few recent works by Alam et al. (2017) and Nguyen et al. (2017b) used visual features only in finding informative images in case of disaster. Rizk et al. (2019) and Mouzannar et al. (2018) used both textual and visual features related to build-infrastructure damage, nature damage, and fire for estimating the damage due to disaster. To the best of our knowledge, no work has been reported where tweet text and images are used together to filter informative tweets from massive social media contents.

The need for a robust disaster-related informative tweet filtering system has motivated us to build a multi-modal system that uses tweet text and images to filter out informative and non-informative posts. The proposed model uses long-short-term-memory (LSTM) for tweet text and convolutional neural network (CNN)–based VGG-16 network for images. We used deep neural network because conventional classification methods require manually engineered features such as TF-IDF vectors, clue words, and Bag-of-Visual-Words. The performance of these conventional classifiers depends heavily on how efficiently the features were extracted. The deep neural network models are better suited for the classification of the disaster-related data than traditional classification approach because they learn features automatically (Nguyen et al., 2016). The proposed model is evaluated on datasets of seven different disasters: Hurricane Harvey, Hurricane Maria, Hurricane Irma, California wildfires, Iran–Iraq earthquake, Mexico earthquake, and Sri Lanka Flood. The contribution of this paper can be summarized below:

1. Development of a multi-modal system to classify informative and non-informative tweets containing either text, image, or both together.
2. Eliminating the need for feature engineering using LSTM and CNN for text and images respectively to extract relevant features.
3. Evaluating the effect of texts and images in the classification of informative contents.
4. Validating the model with cross-event disaster-related datasets to see their efficiency in the early stage of the cross-event disaster.

The rest of the paper is organized as follows: Section 2 discusses the related works; the detailed description of the methodology is discussed in Section 3. The findings of the experimentation are listed in Section 4. Section 5 discusses the overall findings, theoretical contributions, and practical implications of the current work. We conclude the paper in Section 6.

## 2. **Related Literature**

Recently, several works (Imran et al., 2015; Atefeh and Khreich, 2015; Zheng et al., 2018) have been reported for efficiently utilizing the disaster-related social media data for situational awareness. Finding informative contents from the massive social media data is one of the essential tasks for humanitarian organizations. A number of works (Caragea et al., 2016; Nguyen et al., 2016, 2017; Imran et al., 2014; Yu et al., 2019; Ashktorab et al., 2014; Alam et al., 2017; Nguyen et al., 2017; Daly and Thom, 2016) have been reported to identify informative social media contents. However, most of the work focused on the social media text only, whereas images get very little attention in finding informative content. This section is divided into two subsections for better organization of the related literatures: (i) Informative social media text classification and (ii) Informative social media image classification.

### 2.1 Informative social media text classification

The deep neural network–based model is used by Nguyen et al. (2016) to classify messages into informative and not-informative classes. They are further classified informative messages into different classes such as affected individuals, infrastructure and utility damage, and sympathy and support. Caragea et al. (2016) proposed a convolutional neural network–based model to classify tweets into informative and not-informative classes. Their model showed significant improvement over other models that uses n-gram features on flooding datasets. They got their best result of 82.52% in the case of CNN, where they used the Philippines, Colorado, and Queensland floods datasets together as training and Manila floods dataset as testing. Nguyen et al. (2017) proposed a convolutional neural network–based model to classify tweets into informative and not-informative classes. They showed out-of-event data could be considered for training the classifier in the early stage of the events for reducing the effect of the cold-start problem. Caragea et al. (2011) used keyword-based classification and SVM techniques to classify Haiti earthquake tweets into the multi-label setting. They considered several classes such as medical emergency, food shortage, hospital/clinic services in their analysis and achieved F1-scores of 0.47 and 0.59 for the keyword-based classification and SVM, respectively. Imran et al. (2014) developed an Artificial Intelligence for Disaster Response (AIDR) platform to classify Twitter messages into the user-defined classes in real-time. They used human and machine intelligence for labeling a subset of disaster-related messages and trained the model to classify new messages automatically. They tested their platform with the Pakistan earthquake (2013), where they classified messages into informative and not-informative classes with an AUC of 80%. Aipe et al. (2018) developed a deep Convolutional Neural Network (CNN)–based model for multi-label classification of crisis-related tweets. They also explored the uses of Twitter-centric textual features such as hashtags, user-mentions, and keywords extracted from the URLs in the classification task. They found the positive influence of the Twitter-centric features on the performance of the classifier. Their model achieved F1-scores of 0.75 to 0.98 for the seven different categories, such as Casualties and Public Impact, Collateral

Damages, General Awareness, Voluntary Services, Sympathy and Emotion, Crisis-specific Information, and Non-informative. Yu et al. (2019) used CNN, support vector machine (SVM), and logistic regression (LR) to classify tweets related to Hurricane Sandy, Hurricane Harvey, and Hurricane Irma. They classified tweets into different classes such as Caution and Advice, Information Sources, Casualties and Damage, Infrastructure and Resources, and Donation and Aid. They tested their model with two different settings (i) event-specific data and (ii) out-of-event data and achieved F1-score in the range of 0.31 to 0.80. Their CNN-based model performed best in comparison of SVM and LR. Huang and Xiao (2015) manually examined several tweets related to hurricane sandy and code them into different themes. They then used a logistic regression classifier to the tweets to achieve an average F1-score of 0.66. Ashktorab et al. (2014) used several machine-learning techniques such as SVM, logistic regression, Naive Bayes, decision tree, KNN, and supervised latent Dirichlet allocation to identify tweets reporting to damage or casualties. They found their best result in the case of logistic regression with an F1-score of 0.65. Imran et al. (2013b) used informative messages posted during Joplin 2011 and Sandy 2012, and then they used a model based on conditional random fields to extract valuable information from those informative tweets. They achieved the detection rate of 25% to 91% when tested with the event-specific dataset and 1% to 49% when they trained and tested their model with the combination of both Joplin 2011 and Sandy 2012. Imran et al. (2013a) performed three tasks: (i) classified tweets into informative, personal, and others classes; (ii) classified informative tweets into different classes such as Caution, Donation, Casualty, and Information Source; and (iii) extracted several information from the informative tweets such as Location references, Source, and Type of Caution. They used several textual features from the tweet and used the Naive Bayes classifier for both classification task, whereas they used Stanford Named Entity Recognizer to extract information from informative tweets. They got AUC of 0.828 in finding informative tweets and got an F1-score of 0.562 to 0.809 in classifying those informative tweets into further classes. Their information extraction model achieved a precision of 0.47 to 0.93 in finding various information nuggets. Olteanu et al. (2014) created a lexicon of frequently appearing crisis-related terms in the relevant messages. They used this lexicon to automatically identify new terms for a given crisis and query Twitter API to extract crisis-related messages. Graf et al. (2018) extracted linguistic, emotional, and sentimental features from the disaster-related messages and developed a cross-domain classifier. They performed extensive experiments with 26 different disaster-related datasets. They found their best result with an average accuracy of 80% in the case of cross-domain classification, where they used 25 datasets for training and the remaining one for testing. Li et al. (2015) applied the Naive Bayes classifier on the Hurricane Sandy and Boston Marathon bombing Twitter data to study the applicability of domain adaption for mining disaster-related tweets. Rudra et al. (2016) developed a framework to classify Nepal Earthquake tweets into different classes. They summarize those classified tweets to generate comprehensive abstractive summaries. Cameron et al. (2012) developed the Emergency Situation Awareness-Automated Web Text Mining (ESA-AWTM) system that identifies the relevant Twitter messages. Then those relevant messages are used to inform situation awareness of the disaster-related incidents. Verma et al. (2011) build a classifier that used automatically extracted linguistic features to categorize tweets. Their system achieved over

80% in classifying the tweets to contribute the situational awareness. The survey regarding the processing of social media messages and their contributions to situation awareness can be seen in Imran et al. (2015). Some of the potential work which uses social media texts for the classification task are listed in Table 1.

### 2.2 Informative social media image classification

The Image4Act framework is developed by Alam et al. (2017) for identifying relevant images posted on the social media platform to help humanitarian organizations. They tested their framework for the Queensland Australian Cyclone, 2017, and achieved the precision of 0.67 and 0.92 in finding relevant and duplicate images, respectively. Nguyen et al. (2017) developed a pipeline to detect irrelevant and redundant images during a disaster from social media streams. The detection of irrelevant images is done using a transfer-learning approach based on deep neural networks. For the detection of redundant images, they used perceptual hashing techniques. Chaudhuri and Bose (2019) used earthquake-related images and applied the convolutional neural network to identify the human body part from the debris and achieved an accuracy of 83.2%. Daly and Thom (2016) used Flicker images and extracted features from the images to detect the fire event. They found a recall of 91% and a precision of 93% in detecting fire from the images. Lagerstrom et al. (2016) used bush fire-related images of the Australian state of NSW and classified them into the fire and not-fire classes with an accuracy of 86%. Nguyen et al. (2017) used a deep convolutional neural network for classifying disaster-related social media images into severe, mild, and no-damage classes to analyze the impact of the disaster. They used Nepal Earthquake, Ecuador Earthquake, Hurricane Matthew, Typhoon Ruby, and Google Images datasets and trained event-specific as well as cross-event classifier. Their CNN model outperformed Bag-of-Visual-Words (BoVW) techniques and achieved the F1 scores in the range of 0.67 to 0.89.

Recently, researchers have proposed multi-modal systems utilizing the tweet text and images both for finding relevant information from social media. Rizk et al. (2019) developed a multi-modal disaster-related classifier to classify Twitter data into the built-infrastructure damage and nature damage classes. They concatenated semantic features from tweet text and visual features from the image and achieved an accuracy of 92.43%, whereas a model that uses only visual features achieved an accuracy of 91.10%. Mouzannar et al. (2018) developed a multi-modal system based on the deep-learning framework to classify users post into Fire, Floods, Natural landscape damage, Infrastructural damage, Injuries and dead people, and Non-damage classes. They used CNN-based Inception model for image and CNN model for text and combined textual and visual features to classify users' posts and achieved accuracy of 92.62%.

The recently developed multi-modal system is focused on classifying the social media contents into various damage related classes such as build-infrastructure damage, natural damage, and non-damage. None of the works utilized images with the tweet text in finding informative content from the massive social media contents. In this work, we are extracting features from images and combined these features with the features extracted from tweet

text to investigate the role of images in finding informative Twitter contents in the case of the disaster.

**Table 1** List of some potential works for the classification of social media text

| Author | Task | Techniques | Data | Evaluation Metrics | | |
|---|---|---|---|---|---|---|
| | | | | Accuracy | AUC | F1-score |
| Caragea et al. (2016) | Informative vs Not-informative | CNN | Philippines floods (2012), Colorado floods (2013), etc. | 75.90–82.52 | – | – |
| Nguyen et al. (2016) | Informative vs Not-informative and others | CNN | Nepal Earthquake and others | – | 67–78 | – |
| Nguyen et al. (2017) | Informative vs Not-informative | Support vector Machine, Logistic Regression, Random Forest and CNN | Nepal Earthquake, California Earthquake, Cyclone, etc. | – | 50.12–94.17 | – |
| Imran et al. (2014) | Informative vs Not-informative | – | Pakistan Earthquake, 2013 | – | 80 | – |
| Yu et al. (2019) | Caution and Advice, Casualties and Damage, Infrastructure and Resources, etc. | CNN, Support vector machine, Logistic regression | Hurricane Sandy, Hurricane Harvey, and Hurricane Irma | – | – | 0.31–0.80 |
| Huang and Xiao (2015) | Relief, Utility recovery, etc. | Logistic regression | Hurricane Sandy | – | – | 0.00–0.92 |
| Ashktorab et al. (2014) | Damage or casualties vs others | Logistic regression, SVM, KNN, etc. | Christchurch earthquake, Hurricane, Tornado, etc. | 70.0–86.0 | 69–88 | 0.50–0.65 |
| Aipe et al. (2018) | General Awareness, Sympathy and Emotion, Non-informative, etc. | CNN | California Earthquake, Nepal Earthquake, India Flood, etc. | – | – | 0.75–0.98 |
| Caragea et al. (2011) | Medical emergency, food shortage, hospital/clinic services, etc. | Keyword-based classification, SVM | Haiti earthquake | – | – | 0.47–0.59 |

## 3. Methodology

The overall architecture of the proposed multi-modal system is shown in Figure 3. The system consists of two parallel deep neural architectures: (i) long–short-term-memory (LSTM) network for processing textual data and (ii) VGG-16 network for processing images. The tweet text is embedded into a vector form using an Embedding layer shown at the upper left part of Figure 3. This embedded tweet text is then passed through two LSTM layers to extract features from them, which is used to classify the tweet text into informative or not-informative classes. For the image, Convolutional Neural Network (CNN)–based pre-trained VGG-16 model is used to extract features from them. All the weights of the VGG-16 network are marked as non-trainable except the weight between the last dense layer and the output layer. For the multi-modal setting, the feature vector coming from the second last dense layer of VGG-16 is passed through another dense layer containing 256 neurons, as shown in Figure 3. This 256-dimensional feature vector coming from VGG-16 is then concatenated with the 256-dimensional tweet text feature coming from the last LSTM layer to make a 512-dimensional combined feature vector. This 512-dimensional feature vector is then used to predict informative and not-informative Twitter contents. In the following subsections, we will describe the data pre-processing, image classification, text classification, multimodal system, and majority voting scheme: (i) data description and pre-processing, (ii) image classification (VGG-16), (iii) tweet text classification (long-short-term-memory), (iv) multi-modal system (VGG-16 + LSTM), and (v) majority voting.

### 3.1 Data description and pre-processing

The current research uses the dataset published by Alam et al. (2018) to validate the proposed system. It contains seven different disaster-related datasets: (i) Hurricane Harvey, (ii) Hurricane Maria, (iii) Hurricane Irma, (iv) Mexico earthquake, (v) Iran–Iraq earthquake, (vi) California Wildfire, and (vii) Sri Lanka flood. The detailed description regarding the time period and keywords used in the collection for the datasets can be seen in Alam et al. (2018). Here, we are listing the definition of each of the classes mentioned in the datasets which we will use to validate the proposed system: (i) informative: if the tweet/image is useful for humanitarian aid, (ii) not informative: if the tweet/image is not useful for humanitarian aid. Some sample tweets with images for informative and not-informative classes are shown in Figures 1 and 2, respectively. During the creation of the dataset, if a tweet contains more than one image URL, then all the images are downloaded and used the same tweet text with all corresponding images. That means the dataset contains duplicate tweet text in several cases. The data sample containing duplicate tweet text is removed. Finally, we randomly took an equal sample of informative and not-informative tweet text for further processing. The detail description of the datasets is shown in Table 2. In case of tweet texts, symbols such as “#,” “@,” “!,” “&,” and “%” do not contribute to the classification task, so these are removed from the dataset, and all the tweet texts are converted into the lower case. In the case of images, all the images are converted into equal

sizes of $(224 \times 224 \times 3)$. To do the normalization, the pixel matrix of the image is divided by the maximum value, i.e., 255. The normalized matrix is then used by the proposed system to train and test the model. In all the cases, out of the total data sample, 75% of them were used for training, and the remaining 25% sample was used for testing the performance of the models.

**Table 2** Number of informative (Info) and not-informative (Not-info) data samples for different disasters

| | Hurricane Harvey (1940) | | Hurricane Maria (2972) | | Hurricane Irma (1672) | | Mexico Earthquake (624) | | Iran–Iraq Earthquake (168) | | California Wildfires (648) | | Sri Lanka Flood (572) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Info | Not-info | Info | Not-info | Info | Not-info | Info | Not-info | Info | Not-info | Info | Not-info | Info | Not-info |
| Tweet text | 970 | 970 | 1486 | 1486 | 836 | 836 | 312 | 312 | 84 | 84 | 324 | 324 | 286 | 286 |
| Image | 905 | 1035 | 1423 | 1549 | 701 | 971 | 285 | 339 | 64 | 104 | 328 | 320 | 187 | 385 |

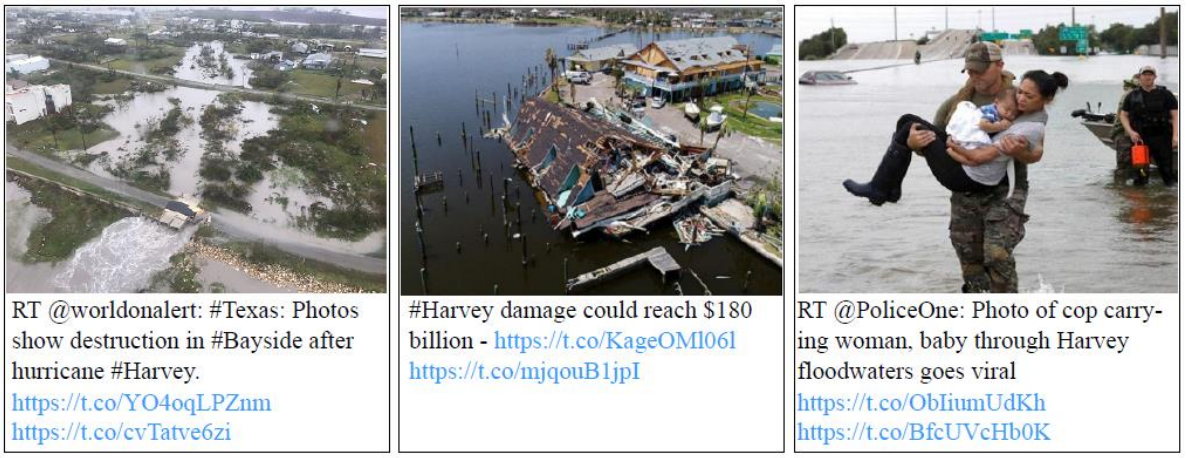


**Fig. 1** Sample informative images with tweets

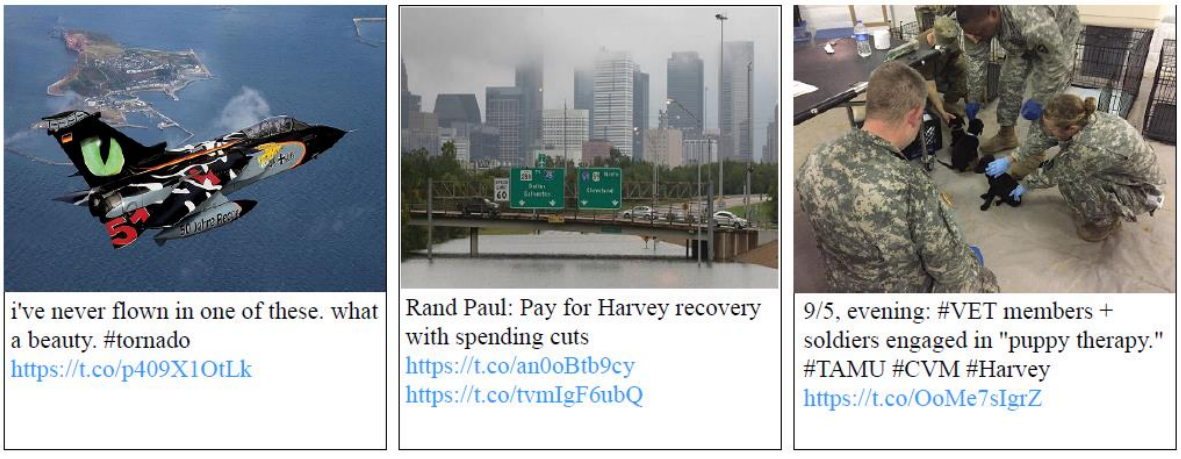


**Fig. 2** Sample not-informative images with tweets

### 3.2 Image classification (VGG-16)

VGG-16 is a deep convolutional neural network architecture designed to classify ImageNet datasets into the 1000 classes. It consists of 13 convolutional layers, followed by three fully connected layers. It takes an image size of $(224 \times 224 \times 3)$ as an input and performs convolution operation using a $(3 \times 3)$ filter. The detailed description regarding the layers and parameters of the VGG-16 network can be seen in Simonyan and Zisserman (2014). The uniform architecture of VGG-16 is very appealing, and currently, it is considered as the most preferred choice for extracting features from images. This VGG-16 network is proved to be effective for the number of image classification tasks (Nguyen et al., 2017; Alam et al., 2017; Nguyen et al., 2017; Simonyan and Zisserman, 2014). Due to the diverse applications of the VGG-16 model for various image processing tasks, the proposed work uses this network. The last layer of the network can be adapted according to the type of classification. As our case is related to binary classification, two neurons are used at the output layer, one for the informative and another one for the not-informative class. The overall architecture of the VGG-16 model can be seen in Figure 3. The weights of the VGG-16 model up to the second dense layer are marked as non-trainable, which is represented in Figure 3 by a dotted box. The weights between the second last dense layer and output layers are trained by passing the image through the network. The model uses softmax activation function with categorical cross-entropy as a loss function, which can be defined by Equations 1 and 2, respectively.

$$\text{Softmax function} = \hat{y}_i = \phi(x_i) = \frac{e^{x_i}}{\sum_{k=1}^{M} e^{x_k}}, \text{where } k = 1,2,\dots,M, \text{and } x_i \in R \qquad (1)$$

$$\text{Categorical cross entropy} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \qquad (2)$$

where $x_i$ is the numerical value coming at the output layer from its previous layer and $M$ represents the number of classes. The softmax function is calculating the probabilities of each target class over all possible target classes. In the second equation, $y_i$ represents one-hot vector for the number of classes and $\hat{y}_i$ represents the predicted class probability of the model for the *i*th training sample in a batch of *N* training sample. In the convolutional layers of VGG-16, Rectified Linear Unit (ReLU) (Nair and Hinton, 2010) is used as an activation function. The ReLU function is defined as: $f(x) = \max(0, x)$, it means for all values of $x < 0$ it returns 0 and for $x > 0$ it returns $x$ itself. The model uses Adam (Kingma and Ba, 2014) as the optimizer. The hyper-parameters used in this study are listed in Table 3. A number of the experiments were carried out to determine the best value of batch size and learning rate, which are found to be 10 and 0.001, respectively.

**Table 3** Hyper-parameter settings for the proposed model

| | LSTM (Tweet text) | VGG-16 (Image) | Multi-modal (Tweet text + Image) |
|---|---|---|---|
| Loss function | Categorical cross entropy | Categorical cross entropy | Categorical cross entropy |
| Optimizer | Adam | Adam | Adam |
| Learning rate | 0.001 | 0.001 | 0.001 |
| Epochs | 200 | 200 | 200 |
| Batch size | 32 | 10 | 10 |
| Activation function | tanh, Softmax | ReLU, Softmax | ReLU, tanh, Softmax |

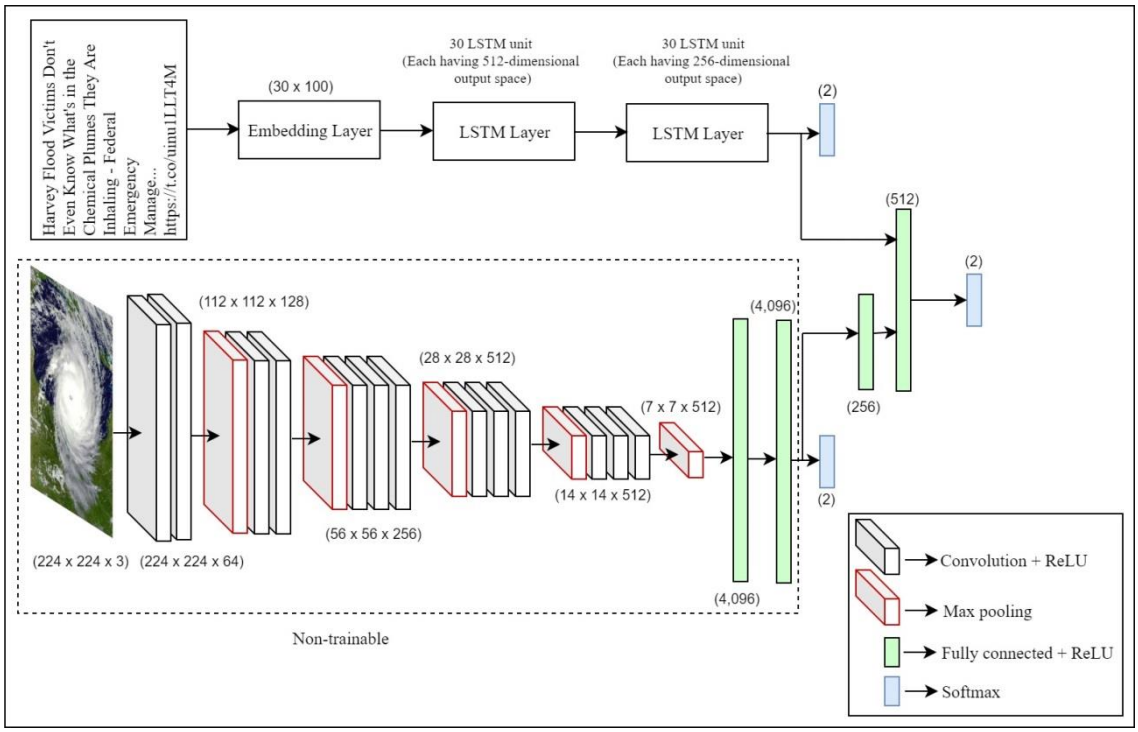


**Fig. 3** Proposed multi-modal neural network model for the classification of Twitter Contents

### 3.3 Tweet text classification (long-short-term-memory)

In this work, the tweet is classified using the long–short-term-memory (LSTM) network. LSTM is designed to remember important information for a longer period of time (Hochreiter and Schmidhuber, 1997; Sundermeyer et al., 2012). As shown in Figure 3, the tweet texts are passed through an embedding layer to get an embedded matrix. This tweet matrix is fed to LSTM layers one after the other. In the first LSTM layer, 30 LSTM units are used, each having 512-dimensional output space. Similarly, in the second LSTM layer, 30 LSTM units are used, each having 256-dimensional output space. The output coming from the second LSTM layer is then connected to 2 neurons, one for informative and other for not-informative classes. The categorical cross-entropy with the softmax activation function is used with Adam optimizer. The model is tested by varying the learning rate and batch size; the best performance was achieved with the learning rate of 0.001 and the batch

size of 32. The detailed hyper-parameter settings for the model are listed in Table 3. The detail description regarding the creation of the tweet matrix and internal architecture of the LSTM unit can be seen in the subsequent sections.

### 3.3.1 Tweet text matrix representation

The word embedding of the tweet is used to feed input to the model. The word embedding represents each word of the corpus into a predefined fixed size real-valued vector. It creates a similar vector for words having similar meanings. The pre-trained word vector GloVe (Global Vectors for word representation) (Pennington et al., 2014) is used as the look-up matrix for this experiment. In our case, 100-dimensional GloVe word vector embedding (glove.twitter.27B.100d.txt)[2] is used, which is trained by Google on 27 billion words of tweets. The advantage of using GloVe is, it reduces the computational overhead of the model. Tweet matrix ($T_i$) can be represented as:

$$T_i = \begin{matrix} W_1 & W_2 & W_3 & \dots & W_m \end{matrix} \begin{bmatrix} e_{11} & e_{21} & e_{31} & \cdots & e_{m1} \\ e_{12} & e_{22} & e_{32} & \dots & e_{m2} \\ e_{13} & e_{23} & e_{33} & \dots & e_{m3} \\ \vdots & \dots & \dots & \dots & \vdots \\ e_{1K} & e_{2K} & e_{3K} & \cdots & e_{mK} \end{bmatrix}$$

where $w_1, w_2, w_3, \dots, w_m$ represents the number of words in a tweet, and $e_{m1}, e_{m2}, e_{m3}, \dots, e_{mK}$ represents the embedding of the word $W_m$. This tweet matrix has $(K \times m)$ dimension, where $K$ is the dimension of the embedding vector and $m$ is the number of words in the tweet. In this work, we fixed the total number of words for a tweet to 30, as most of the tweets contain 30 or less than 30 words. Padding is used where it is required to make all the tweets into the same length. As 100-dimensional GloVe embedding is used, so in our case, the dimension of a tweet matrix is $(30 \times 100)$, which is represented in Figure 3. This tweet matrix is then used by LSTM layers to learn the salient features from them. As the tweet matrix is represented in $(30 \times 100)$ dimension, 30 LSTM units are used to process the embedding of each word. The detail description of an LSTM unit can be seen in Section 3.3.2.

### 3.3.2 Long–Short Term Memory (LSTM) unit

This section discusses the detail working principle of an LSTM unit. Each LSTM unit contains four components: (i) forget gate ($f_t$), (ii) input gate ($i_t$), (iii) cell state ($C_t$), and (iv) output gate ($O_t$). The cell state keeps the relevant information throughout the processing of the sequences. This cell state can be considered as the "memory" of the network. The information is added or deleted using gates throughout the journey of the cell state. The forget gate decides which information should be kept or thrown away based on their importance. Input gate is used to update the cell state, and the output gate decides what

[2] It is freely available at https://nlp.stanford.edu/projects/glove/

will be the next hidden state. During training the model, these gates learn which information in a sequence is important to keep or forget. They pass only those information to the cell state, which is important for the prediction. The detailed internal architecture of an LSTM unit is shown in Figure 4. In the figure, $C_{t-1}$ and $C_t$ represent cell state for time step $t-1$ and $t$, respectively. Similarly, $h_{t-1}$ and $h_t$ represents hidden layer output at time step $t-1$ and $t$, respectively. The input feature to the LSTM unit is denoted by $X_t$. It contains sigmoid and tanh activation function which can be defined by Equations 3 and 4, respectively.

Sigmoid function: $\sigma(x_i) = \frac{1}{1+e^{-x_i}}$ , where $x_i \in R$ (3)

$\tanh(x_i) = \frac{e^{x_i}-e^{-x_i}}{e^{x_i}+e^{-x_i}}$ , where $x_i \in R$ (4)

The value of sigmoid and tanh activation function ranges from 0 to 1 and –1 to 1 respectively. These values are basically responsible for all the gate operation. The forget gate $(f_t)$, input gate $(i_t)$ , cell state $(C_t)$, and output gate $(O_t)$ mathematically can be represented by Equations 5, 6, 7, and 8, respectively.

Forget gate $(f_t) = \sigma(\alpha_f . [h_{t-1}] + \beta_f)$ (5)

Input gate $(i_t) = \sigma(\alpha_i \cdot [h_{t-1}, x_t] + \beta_i)$ (6)

Cell state $(C_t) = f_t \times C_{t-1} + i_t \times C'_t$ (7)

$C'_t = \tanh(\alpha_C \cdot [h_{t-1}, x_t] + \beta_C)$

Output gate $(O_t) = \sigma(\alpha_O \cdot [h_{t-1}, x_t] + \beta_O)$ (8)

where $\alpha_f, \alpha_i, \alpha_C, \alpha_O$ are the weight matrices and $\beta_f, \beta_i, \beta_C, \beta_O$ are the bias values for the forget gate, input gate, cell state, and output gate, respectively. Finally, the hidden layer output at time step $t$ can be defined as $h_t = O_t \times \tanh(C_t)$. This hidden layer output is then connected with the next LSTM unit.

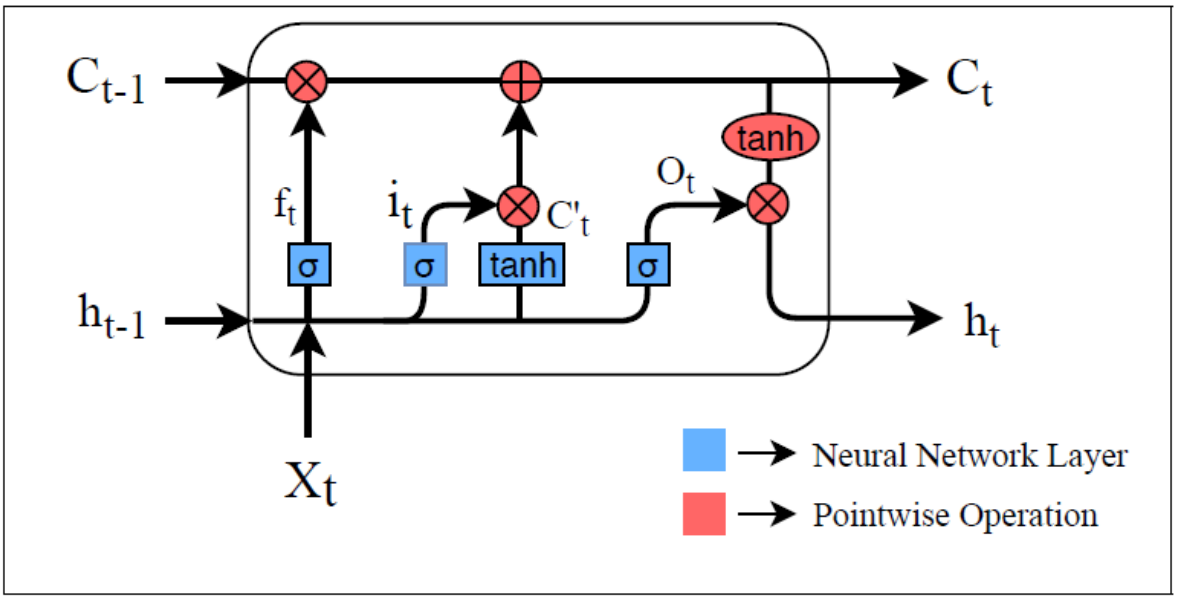


**Fig. 4** The detailed internal architecture of a LSTM unit

### 3.4 Multi-modal system (VGG-16 + LSTM)

The proposed multi-modal system uses both tweet texts and images for the classification of informative and not-informative content. The second-last layer of the VGG-16 model containing 4,096 neurons in their dense layer is mapped to another dense layer containing 256 neurons. The last layer of the LSTM model is then concatenated with the 256-dimensional feature map of the VGG-16 model. Total of 512-dimensional feature map is generated by concatenating both the layers as can be seen from Figure 3. Finally, concatenated feature maps are then mapped to the output layer containing two neurons, one for the informative and one for the not-informative classes. The label of tweet text is used as the final label for the concatenated features of tweet texts and images. As in the case of the VGG-16 model, the weights are marked as non-trainable up to the second last layer, so here, the same procedure is applied. Similarly, softmax activation function at the output layer, categorical cross-entropy as a loss function, Adam, as the optimizer with a learning rate of 0.001, is used. The model performed best with the learning rate and batch size of 0.001 and 10, respectively.

### 3.5 Majority voting

In the majority voting strategy, all the three models LSTM (Tweet text), VGG-16 (Image), and Multi-modal (Tweet text + Image) are used. The prediction of each of the models is considered, and the final prediction is assigned based on the majority. If at least two models predicted a class, then the final label is assigned with that class label. The finding of this strategy is discussed in detail in Section 4.

## 4. Results

The extensive experiments have been done to validate the proposed model under two categories: (i) event specific experiment: the system was trained and tested with the same event data only and (ii) cross-event experiment: the system was trained with a specific event but tested with other event data. The performance of the system has been evaluated using precision, recall, and F1-score. The description of the used evaluation metrics is given in Section 4.1.

### 4.1 Evaluation metrics

– Precision: Precision for a class (say informative class) can be defined as, number of accurately predicted informative contents to the total number of predicted informative contents. The value of precision varies between 0 and 1, where 0 means the worst performance and 1 means the best performance.

$$\text{Precision} = \frac{\text{Number of accurately predicted informative contents}}{\text{Total number of predicted informative contents}}$$

– Recall: Recall for a class (say informative class) can be defined as, number of accurately predicted informative contents to the total number of informative contents in the dataset. The value of recall varies from 0 to 1, where 0 is the worst performance and 1 is the best performance.

$$\text{Recall} = \frac{\text{Number of accurately predicted informative contents}}{\text{Total number of actual informative contents}}$$

– F1-score: F1-score is the harmonic mean between precision and recall. The value of F1-score varies from 0 to 1, where 0 indicates the worst performance, whereas 1 indicates the best performance.

$$\text{F1} - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The weighted average of precision, recall, and F1-score of informative and not-informative classes are reported to evaluate the performance of the proposed model. In our experiment, all the three models LSTM (Tweet text), VGG-16 (Image), and Multi-modal (Tweet text + Image) are trained separately with the dataset of (i) Hurricane Harvey, (ii) Hurricane Irma, (iii) Hurricane Maria, (iv) Iran–Iraq earthquake, (v) Mexico earthquake, (vi) California wildfires, and (vii) Sri Lanka flood. Then, all the trained models are tested individually with all the possible combinations of testing data. For example, if the LSTM (Tweet text) is trained with say Hurricane Harvey, it is tested with all seven datasets. Similarly, when LSTM (Tweet text) is trained with Hurricane Irma, it is tested with all seven datasets. Likewise, for one model, say LSTM (Tweet text), we performed 49 testing experiments. We

have three models, LSTM (Tweet text), VGG-16 (Image), and Multi-modal, and for each, we formed 49 test cases. Therefore, in total, 147 test cases were formed to evaluate the performance of the models. The results for the models when trained with (i) Hurricane Harvey, (ii) Hurricane Irma, (iii) Hurricane Maria, (iv) Iran–Iraq earthquake, (v) Mexico earthquake, (vi) California wildfires, and (vii) Sri Lanka flood separately and tested with all the possible combinations of the testing data are shown in Tables 4, 5, 6, 7, 8, 9, and 10 respectively. The confusion matrices of the best performing model when tested with the same event data for (i) Hurricane Harvey, (ii) Hurricane Irma, (iii) Hurricane Maria, (iv) Iran–Iraq earthquake, (v) Mexico earthquake, (vi) California wildfires, and (vii) Sri Lanka flood are shown in Figures 5, 6, 7, 8, 9, 10, and 11, respectively.

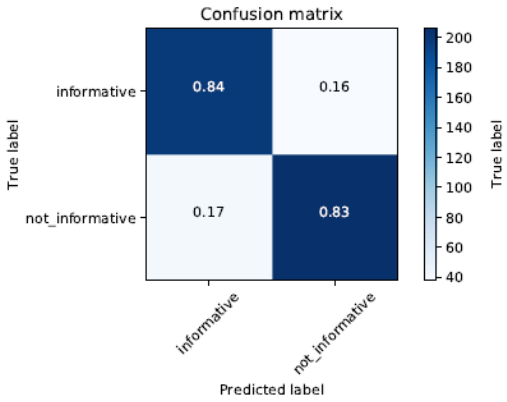


**Fig. 5** Confusion matrix when system was trained and tested with Hurricane Harvey dataset

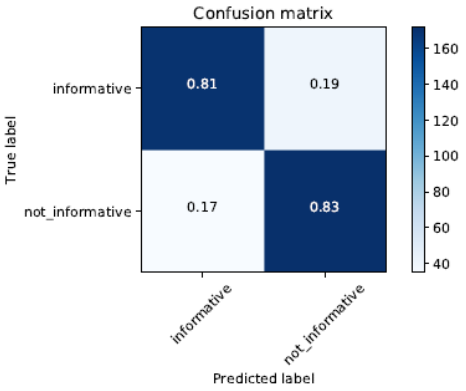


**Fig. 6** Confusion matrix when system was trained and tested with Hurricane Irma dataset

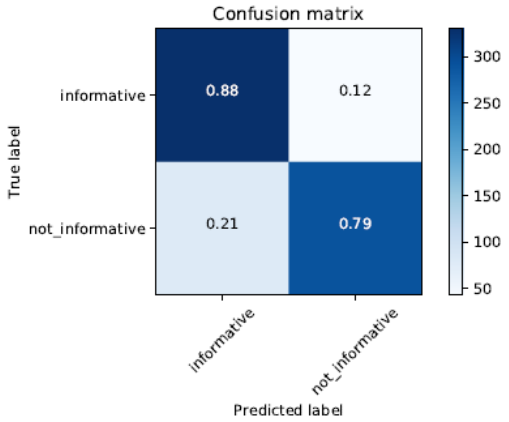


**Fig. 7** Confusion matrix when system was trained and tested with Hurricane Maria dataset

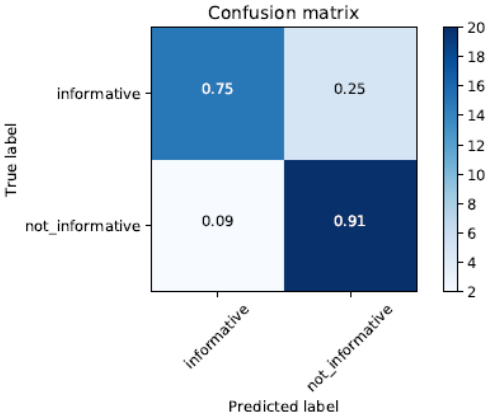


**Fig. 8** Confusion matrix when system was trained and tested with Iraq–Iran earthquake dataset

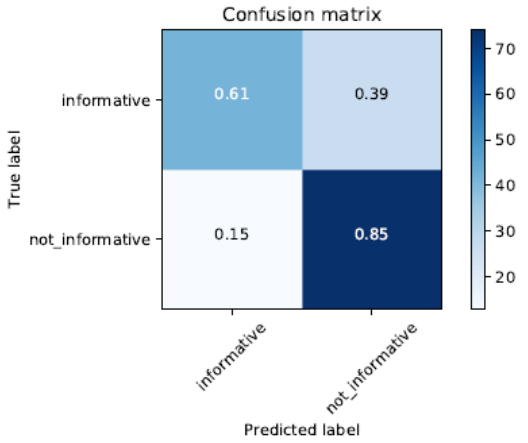


**Fig. 9** Confusion matrix when system was trained and tested with Mexico earthquake dataset

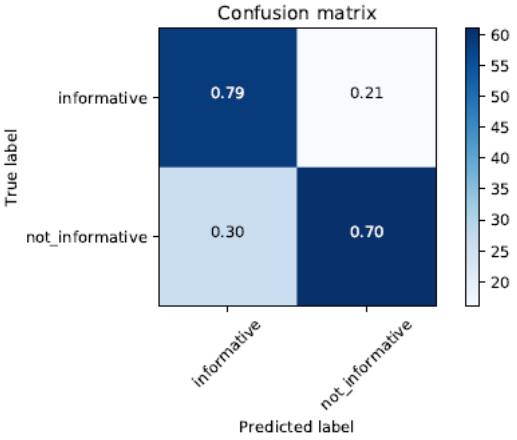


**Fig. 10** Confusion matrix when system was trained and tested with California wildfire dataset

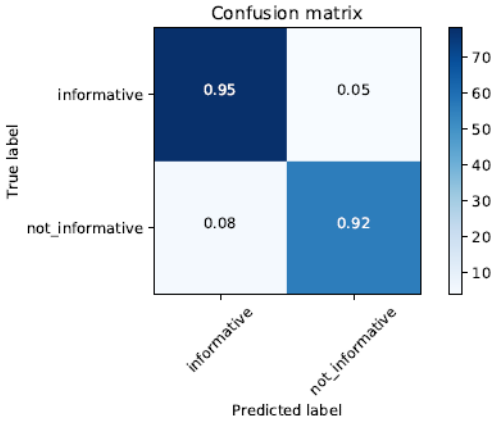


**Fig. 11** Confusion matrix when system was trained and tested with Sri Lanka flood dataset

**Table 4** Results of various models when it is trained with Hurricane Harvey dataset

| Hurricane Harvey | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Hurricane Harvey | | | Hurricane Maria | | | Hurricane Irma | | | California Wildfires | | | Mexico Earthquake | | | Iraq–Iran Earthquake | | | Sri Lanka Flood | | |
| Models | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| LSTM(Tweet Text) | 0.81 | 0.81 | 0.81 | 0.74 | 0.72 | 0.72 | 0.74 | 0.70 | 0.69 | 0.56 | 0.56 | 0.56 | 0.65 | 0.65 | 0.65 | 0.52 | 0.52 | 0.52 | 0.84 | 0.80 | 0.80 |
| VGG-16 (Image) | 0.81 | 0.81 | 0.81 | 0.70 | 0.71 | 0.70 | 0.74 | 0.74 | 0.74 | 0.75 | 0.75 | 0.75 | 0.82 | 0.81 | 0.81 | 0.84 | 0.83 | 0.83 | 0.79 | 0.76 | 0.77 |
| Multi-modal | 0.83 | 0.82 | 0.82 | 0.76 | 0.76 | 0.76 | 0.75 | 0.75 | 0.75 | 0.63 | 0.62 | 0.62 | 0.72 | 0.66 | 0.65 | 0.66 | 0.62 | 0.60 | 0.87 | 0.87 | 0.87 |
| Majority voting | 0.84 | 0.84 | 0.84 | 0.76 | 0.76 | 0.76 | 0.76 | 0.75 | 0.75 | 0.63 | 0.63 | 0.63 | 0.70 | 0.69 | 0.69 | 0.60 | 0.60 | 0.59 | 0.87 | 0.85 | 0.85 |

**Table 5** Results of various models when it is trained with Hurricane Irma dataset

| Hurricane Irma | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Hurricane Irma | | | Hurricane Harvey | | | Hurricane Maria | | | Sri Lanka flood | | | Mexico Earthquake | | | Iraq–Iran Earthquake | | | California Wildfires | | |
| Models | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| LSTM(Tweet Text) | 0.83 | 0.83 | 0.83 | 0.74 | 0.72 | 0.71 | 0.76 | 0.76 | 0.75 | 0.84 | 0.81 | 0.81 | 0.69 | 0.63 | 0.62 | 0.71 | 0.64 | 0.62 | 0.68 | 0.65 | 0.65 |
| VGG-16 (Image) | 0.83 | 0.83 | 0.83 | 0.74 | 0.74 | 0.74 | 0.76 | 0.76 | 0.76 | 0.76 | 0.76 | 0.76 | 0.83 | 0.83 | 0.83 | 0.74 | 0.71 | 0.72 | 0.72 | 0.71 | 0.70 |
| Multi-modal | 0.81 | 0.81 | 0.81 | 0.76 | 0.75 | 0.75 | 0.78 | 0.78 | 0.78 | 0.83 | 0.78 | 0.77 | 0.64 | 0.61 | 0.60 | 0.68 | 0.64 | 0.63 | 0.66 | 0.66 | 0.66 |
| Majority voting | 0.82 | 0.82 | 0.82 | 0.77 | 0.75 | 0.74 | 0.79 | 0.79 | 0.79 | 0.82 | 0.78 | 0.78 | 0.71 | 0.67 | 0.67 | 0.74 | 0.69 | 0.68 | 0.70 | 0.70 | 0.70 |

**Table 6** Results of various models when it is trained with Hurricane Maria dataset

| Hurricane Maria | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Hurricane Maria | | | Hurricane Harvey | | | Hurricane Irma | | | Sri Lanka flood | | | Mexico Earthquake | | | Iraq–Iran Earthquake | | | California Wildfires | | |
| Models | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| LSTM(Tweet Text) | 0.83 | 0.83 | 0.82 | 0.72 | 0.72 | 0.72 | 0.77 | 0.76 | 0.76 | 0.90 | 0.90 | 0.90 | 0.68 | 0.57 | 0.53 | 0.77 | 0.57 | 0.49 | 0.61 | 0.60 | 0.60 |
| VGG-16 (Image) | 0.78 | 0.78 | 0.78 | 0.73 | 0.73 | 0.73 | 0.75 | 0.75 | 0.75 | 0.81 | 0.81 | 0.81 | 0.79 | 0.79 | 0.79 | 0.82 | 0.81 | 0.81 | 0.67 | 0.65 | 0.63 |
| Multi-modal | 0.83 | 0.83 | 0.83 | 0.77 | 0.76 | 0.76 | 0.78 | 0.72 | 0.71 | 0.90 | 0.90 | 0.90 | 0.66 | 0.58 | 0.55 | 0.77 | 0.55 | 0.45 | 0.60 | 0.59 | 0.59 |
| Majority voting | 0.84 | 0.84 | 0.84 | 0.77 | 0.77 | 0.77 | 0.78 | 0.76 | 0.76 | 0.91 | 0.91 | 0.91 | 0.66 | 0.58 | 0.56 | 0.77 | 0.57 | 0.49 | 0.63 | 0.62 | 0.62 |

**Table 7** Results of various models when it is trained with Iraq–Iran earthquake dataset

| Iraq–Iran Earthquake | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Iraq–Iran Earthquake | | | Mexico Earthquake | | | Hurricane Harvey | | | Hurricane Maria | | | Hurricane Irma | | | California Wildfires | | | Sri Lanka flood | | |
| Models | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| LSTM(Tweet Text) | 0.81 | 0.81 | 0.81 | 0.66 | 0.67 | 0.66 | 0.59 | 0.51 | 0.35 | 0.75 | 0.50 | 0.34 | 0.62 | 0.50 | 0.34 | 0.75 | 0.55 | 0.40 | 0.76 | 0.44 | 0.28 |
| VGG-16 (Image) | 0.86 | 0.86 | 0.85 | 0.81 | 0.79 | 0.78 | 0.60 | 0.59 | 0.51 | 0.67 | 0.60 | 0.53 | 0.58 | 0.59 | 0.51 | 0.47 | 0.51 | 0.38 | 0.71 | 0.71 | 0.68 |
| Multi-modal | 0.79 | 0.79 | 0.79 | 0.74 | 0.74 | 0.74 | 0.65 | 0.62 | 0.59 | 0.59 | 0.57 | 0.53 | 0.58 | 0.56 | 0.54 | 0.56 | 0.56 | 0.48 | 0.70 | 0.60 | 0.58 |
| Majority voting | 0.84 | 0.83 | 0.83 | 0.75 | 0.75 | 0.75 | 0.68 | 0.55 | 0.44 | 0.56 | 0.51 | 0.41 | 0.54 | 0.51 | 0.40 | 0.66 | 0.56 | 0.42 | 0.78 | 0.52 | 0.44 |

**Table 8** Results of various models when it is trained with Mexico earthquake dataset

| Mexico Earthquake | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mexico Earthquake | | | Iraq–Iran Earthquake | | | Hurricane Harvey | | | Hurricane Maria | | | Hurricane Irma | | | Sri Lanka flood | | | California Wildfires | | |
| Models | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| LSTM(Tweet Text) | 0.71 | 0.71 | 0.71 | 0.63 | 0.60 | 0.58 | 0.72 | 0.66 | 0.64 | 0.60 | 0.56 | 0.52 | 0.64 | 0.55 | 0.48 | 0.74 | 0.57 | 0.53 | 0.53 | 0.54 | 0.52 |
| VGG-16 (Image) | 0.87 | 0.87 | 0.87 | 0.78 | 0.79 | 0.78 | 0.66 | 0.66 | 0.66 | 0.66 | 0.66 | 0.65 | 0.68 | 0.68 | 0.67 | 0.81 | 0.81 | 0.81 | 0.73 | 0.72 | 0.71 |
| Multi-modal | 0.73 | 0.72 | 0.72 | 0.80 | 0.67 | 0.62 | 0.66 | 0.58 | 0.51 | 0.60 | 0.55 | 0.48 | 0.62 | 0.56 | 0.50 | 0.65 | 0.55 | 0.51 | 0.63 | 0.61 | 0.58 |
| Majority voting | 0.75 | 0.74 | 0.74 | 0.71 | 0.67 | 0.64 | 0.70 | 0.64 | 0.60 | 0.61 | 0.56 | 0.51 | 0.65 | 0.58 | 0.53 | 0.69 | 0.58 | 0.55 | 0.63 | 0.62 | 0.59 |

**Table 9** Results of various models when it is trained with California wildfires dataset

| California Wildfires | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | California Wildfires | | | Hurricane Harvey | | | Hurricane Irma | | | Hurricane Maria | | | Mexico Earthquake | | | Iraq–Iran Earthquake | | | Sri Lanka flood | | |
| Models | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| LSTM(Tweet Text) | 0.64 | 0.64 | 0.64 | 0.45 | 0.47 | 0.41 | 0.51 | 0.50 | 0.45 | 0.50 | 0.50 | 0.46 | 0.42 | 0.47 | 0.42 | 0.65 | 0.64 | 0.63 | 0.43 | 0.41 | 0.39 |
| VGG-16 (Image) | 0.83 | 0.82 | 0.82 | 0.71 | 0.71 | 0.70 | 0.73 | 0.74 | 0.73 | 0.69 | 0.69 | 0.68 | 0.76 | 0.76 | 0.76 | 0.73 | 0.74 | 0.74 | 0.79 | 0.79 | 0.79 |
| Multi-modal | 0.74 | 0.73 | 0.73 | 0.64 | 0.60 | 0.57 | 0.61 | 0.59 | 0.57 | 0.59 | 0.57 | 0.55 | 0.65 | 0.63 | 0.60 | 0.61 | 0.57 | 0.51 | 0.66 | 0.57 | 0.55 |
| Majority voting | 0.75 | 0.74 | 0.74 | 0.63 | 0.60 | 0.58 | 0.60 | 0.59 | 0.57 | 0.62 | 0.59 | 0.57 | 0.66 | 0.65 | 0.62 | 0.62 | 0.60 | 0.56 | 0.66 | 0.58 | 0.56 |

**Table 10** Results of various models when it is trained with Sri Lanka flood dataset

| Sri Lanka flood | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Sri Lanka flood | | | Hurricane Harvey | | | Hurricane Maria | | | Hurricane Irma | | | Mexico Earthquake | | | Iraq–Iran Earthquake | | | California Wildfires | | |
| Models | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| LSTM(Tweet Text) | 0.92 | 0.92 | 0.92 | 0.73 | 0.70 | 0.69 | 0.74 | 0.69 | 0.68 | 0.76 | 0.72 | 0.71 | 0.61 | 0.62 | 0.60 | 0.70 | 0.67 | 0.66 | 0.58 | 0.57 | 0.53 |
| VGG-16 (Image) | 0.87 | 0.85 | 0.84 | 0.67 | 0.61 | 0.53 | 0.71 | 0.62 | 0.56 | 0.67 | 0.64 | 0.58 | 0.81 | 0.80 | 0.79 | 0.79 | 0.79 | 0.76 | 0.66 | 0.56 | 0.45 |
| Multi-modal | 0.94 | 0.94 | 0.94 | 0.74 | 0.74 | 0.74 | 0.74 | 0.74 | 0.74 | 0.80 | 0.80 | 0.80 | 0.57 | 0.57 | 0.57 | 0.77 | 0.69 | 0.67 | 0.54 | 0.54 | 0.53 |
| Majority voting | 0.93 | 0.93 | 0.93 | 0.75 | 0.72 | 0.71 | 0.74 | 0.70 | 0.68 | 0.75 | 0.72 | 0.71 | 0.63 | 0.63 | 0.62 | 0.72 | 0.69 | 0.68 | 0.58 | 0.57 | 0.53 |

**Table 11** Comparison of the proposed work with the existing methodologies

| | Hurricane Harvey F1-score | Hurricane Maria F1-score | Hurricane Irma F1-score | Iraq–Iran earthquake F1-score | Mexico earthquake F1-score | Sri Lanka Flood F1-score | California Wildfire F1-score |
|---|---|---|---|---|---|---|---|
| SVM (Text) | 0.32 | 0.33 | 0.33 | 0.31 | 0.27 | 0.26 | 0.29 |
| Random Forest (Text) | 0.75 | 0.81 | 0.72 | 0.69 | 0.68 | 0.83 | 0.60 |
| Logistic Regression (Text) | 0.78 | 0.82 | 0.80 | 0.74 | 0.72 | 0.82 | 0.60 |
| CNN (Text) | 0.81 | 0.82 | 0.80 | 0.74 | 0.68 | 0.92 | 0.61 |
| LSTM (Text) | 0.81 | 0.82 | 0.83 | 0.81 | 0.71 | 0.92 | 0.64 |
| Majority voting (Image + Text) | 0.84 | 0.84 | 0.82 | 0.83 | 0.74 | 0.93 | 0.74 |

## 5. Discussion

Our findings suggest that the identification of informative Twitter contents related to disaster using images and text together with a majority voting scheme is better than the models utilizing text alone. It is also found that in the case of earthquakes and wildfires event, images alone are performing better than text in identifying informative contents, but the authenticity of images are questionable as people post old images of similar events during the current disaster event (Gupta et al., 2013; Imran et al., 2015). Therefore, for the experimentation, we have considered the label of text as the final label for the multi-modal settings. The finding of this research also suggests that the embedding of images with the tweet text performed significantly better in identifying disaster-related informative contents. The proposed systems are validated for two different settings: (i) In-event validation and (ii) Cross-event validation. In the case of In-event validation, the system is trained and tested with the same event dataset, whereas in the case of Cross-event validation, the system is trained with one dataset and tested with different event datasets. In both the In-event and Cross-event settings, out of the total 49 sets of testing combinations, the model with the Majority voting scheme outperformed the LSTM model in 39 cases where only tweet texts have been used. In the remaining 10 cases, the Majority voting scheme gives comparable results with respect to the LSTM model.

In-event validation: (a) Hurricane Harvey: The Majority voting scheme and the multi-modal system performed better than the LSTM model where only tweet text was used. The Majority voting scheme achieved an F1-score of 0.84, whereas the LSTM model achieved an F1-score of 0.81 in classifying informative and not-informative Twitter contents. The confusion matrix shown in Figure 5 for the Majority voting system when it is trained and tested with Hurricane Harvey shows that out of 100 informative contents, the model predicts 84 contents as the informative. (b) Hurricane Irma: The best result was obtained for LSTM model with tweet text (F1-score = 0.83) but the majority voting scheme has also given comparable results (F1-score = 0.82). (c) Hurricane Maria: The Majority voting and multi-modal system performed better than the LSTM model, where only the tweet text was used. The Majority voting and Multi-modal system achieved an F1-score of 0.84 and 0.83, respectively, whereas the LSTM model achieved an F1-score of 0.82. (d) Iraq–Iran earthquake: The Majority voting system achieved an F1-score of 0.83, which is better than the LSTM model as it achieved an F1-score of 0.81. (e) Mexico earthquake: The Majority voting achieved an F1-score of 0.74, which is 3% higher than the LSTM model where only tweet text was used. (f) California Wildfire: In this case, the Majority voting system performed better than the LSTM model with a margin of 10% in the F1-score. The Majority voting system achieved an F1-score of 0.74, whereas the LSTM model achieved an F1-score of 0.64. (g) Sri Lanka floods: A similar kind of result is also found in the case of Sri Lanka floods. The Majority voting scheme achieved an F1-score of 0.93, which is 1% higher than the LSTM model. In the case of In-event validation, out of the total seven different event datasets, the combination of text and images both for Majority voting system performed better than the LSTM model for six events, namely, Hurricane Harvey, Hurricane Maria, Iraq–Iran earthquake, Mexico earthquake, California wildfires, and Sri Lanka floods. In the case of Hurricane Irma, the Majority voting system achieved a comparable result to the LSTM model.

Cross-event validation: (a) Hurricane Harvey: When the models are tested with Hurricane Irma and Hurricane Maria, the Majority voting and Multi-modal system performed better than the LSTM model. As the nature of the Hurricane Irma and Hurricane Maria events are

same as the Hurricane Harvey, the proposed model performed significantly well. The Majority voting system achieved an F1-score of 0.76 and 0.75 for Hurricane Maria and Hurricane Irma events, respectively. When the models are tested with the cross-event dataset (California wildfires, Mexico earthquake, and Iraq–Iran earthquake), the performance of the model has been degraded as the nature of the event is changed, although the Majority voting scheme gives a better result in comparison to LSTM. While testing the model with the Sri Lanka flood dataset, the model performed well with an F1-score of 0.85, because the text and images in both the cases contained water as a component. One of the examples for Hurricane Harvey is: "*RT NickABC13: Yes, that's a Cadillac stuck in water. The driver had to be rescued. #Harvey https://t.co/c3c8lv0MQo*", and one example for the Sri Lanka flood is: "*Mora Impact: Port city sees heavy rain, waterlogging #TISNews Click Link- https://t.co/mn9pvB4s1t https://t.co/6ywgXiVJEO*". In both the tweets, people are talking about water, which is one of the possible reasons why these types of cross-event testing perform better. (b) Hurricane Irma: A similar kind of result is found in the case of Hurricane Irma as well. The Majority voting system performed significantly better than the LSTM model in the case of Hurricane Harvey, Hurricane Maria, Mexico earthquake, Iran–Iraq earthquake, and California wildfires events. But in the case of Sri Lanka flood, the F1-score of Majority voting was 0.78, which is less than the LSTM model (F1-score = 0.81). The performance of the multi-modal system is also slightly degraded in comparison to the LSTM model. (c) Hurricane Maria: When the model is trained with Hurricane Maria, the Majority voting system performed better than the LSTM model throughout all the cross-event testing (see Table 6).

(d) Iraq–Iran earthquake: The cross-event testing with the Majority voting system performed significantly better than the LSTM model throughout all the testing events (see Table 7). The Mexico earthquake has the same nature as the Iran–Iraq earthquake. The performance of the Majority voting scheme is better in comparison to the LSTM model by a margin of 9% (F1-score = 0.66 in case of only text, whereas F1-score of 0.75 in case of the Majority voting scheme). (e) Mexico earthquake: Similarly, the system trained with the Mexico earthquake, the Majority scheme performed better than the LSTM model in case of Iraq–Iran earthquake, Hurricane Irma, Sri Lanka flood, and California wildfire (see Table 8). (f) California wildfires: The cross-event testing with the Majority voting system performed significantly well for Hurricane Harvey, Hurricane Irma, Hurricane Maria, Mexico earthquake, and Sri Lanka floods. In the case of the Iraq–Iran earthquake event, the performance of the Majority voting system degraded slightly. (g) Sri Lanka floods: The Majority voting system performed better than the LSTM model throughout all the testing cases (see Table 10). Similarly, the model performed well when it was tested with Hurricane Harvey, Hurricane Irma, and Hurricane Maria in comparison with other cross-event testing cases. The possible reason for this is similar to the case of Hurricane as it contains tweets related to water-logging and area in the image filled with water.

Our results are better than recently proposed similar works by Nguyen et al. (2016, 2017) and Caragea et al. (2016). Nguyen et al. (2016, 2017) used Convolutional Neural Network (CNN), whereas Caragea et al. (2016) used Support Vector Machine (SVM), Random Forest, Logistic Regression, and (CNN) techniques. In order to compare our models with them, we tested these models with the datasets we have used. For SVM, Random Forest, and Logistic Regression unigram, bigram, and trigram TF-IDF features were used. For

CNN, 2-gram, 3-gram, and 4-gram filters were used to extract features from the tweet texts. The result of each of the models is shown in Table 11. The proposed Majority voting scheme outperformed all the existing works across all the datasets except Hurricane Irma. Even LSTM model with tweet text only also outperformed all the text classification techniques such as SVM, Random Forest, Logistic Regression, and Convolutional Neural Network (CNN). One of the limitations of this work is that we have not checked the authenticity of images used for the experimentation. It has been observed that several old and similar images are posted by the users during disaster. Therefore, a system can be developed to filter these old and similar images.

### 5.1 Theoretical contributions

One of the major theoretical contributions of this research is the development of a parallel system with LSTM for tweet text and VGG-16 for images of disastrous scenarios to classify informative and not-informative Twitter contents. The proposed system does not require any human efforts to extract features for training the model. The other theoretical contribution is that the proposed system uses a pre-trained VGG-16 network, which reduces the overall training time of the system as it is required in the case of disaster.

The proposed system is validated with both In-event and Cross-event disasters, and it significantly performed better than the systems where only tweet text was used. Therefore, this system can be better utilized in the situation of cross-event disaster in the early stage where a smaller number of the disaster-specific labeled data is available. The proposed system can be utilized in all the three types of input data, such as only tweet text, only images, and tweet text and images together to classify informative and non-informative Twitter contents.

### 5.2 Implications for practice

The model can be implemented in any system to segregate the informative tweets from non-informative tweets using either text, images, or both together. These informative tweets can then be used by humanitarian organizations to know the floor reality of the disaster. This system can be integrated with any social media platform to filter informative contents form massive social media content. An android application can also be made where this system can separate disaster-related informative contents from the live streaming of social media posts to help people become more situationally aware in the case of disaster. This multi-modal system can be utilized in finding relevant information in other domains also such as finding relevant content related to road accidents and civil unrest if domain-specific training is done.

## 6. Conclusion

The identification of disaster-related informative messages from Twitter is a challenging task as tweets have several grammatical mistakes, non-standard abbreviations, and limited

word space. In this work, a multi-modal system is proposed which utilizes tweet texts as well as images to identify informative Twitter contents. The system uses LSTM and VGG-16 for tweet text and image, respectively. We have used seven different disasters related Twitter datasets and achieved an F1-score of 0.84, 0.84, 0.82, 0.83, 0.74, 0.93, and 0.74 for Hurricane Harvey, Hurricane Maria, Hurricane Irma, Iraq–Iran earthquake, Mexico earthquake, Sri Lanka flood, and California wildfires, respectively, in the case of Majority voting scheme. These results have outperformed the other models where only tweet text is used. This system can also be utilized in other similar kinds of crisis events, as in our case, we have tested it with Hurricane and flood that has achieved significant results. This model can be used for the primary filtration of informative tweets from the massive number of tweets. Then, the informative tweets can be further classified into several classes such as infrastructure and utility damage, affected individuals, injured or dead people, and vehicle damage for providing better rescue and relief operation. The limitation of this work is that we have considered English language tweets only, but during an emergency, people also post their tweets in regional languages. So, a deep neural network–based model can be developed to deal with the issues of multi-linguality. As the F1-score of the proposed approaches varies in the range of 0.74 to 0.93, the system can be enhanced to achieve better accuracy.

**References**


Aipe, A., Ekbal, A., Mukuntha, N., & Kurohashi, S. (2018). Linguistic feature assisted deep learning approach towards multi-label classification of crisis related tweets. In *Proceedings of the 15th ISCRAM Conference* (pp. 705–717).

Akter, S., & Wamba, S. F. (2017). Big data and disaster management: a systematic review and agenda for future research. *Annals of Operations Research*, https://doi.org/10.1007/s10479-017-2584-2.

Alam, F., Imran, M., & Ofli, F. (2017). Image4act: Online social media image processing for disaster response. *In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017* (pp. 601–604). ACM.

Alam, F., Ofli, F., & Imran, M. (2018). Crisismmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM).*

Ashktorab, Z., Brown, C., Nandi, M., & Culotta, A. (2014). Tweedr: Mining twitter to inform disaster response. In *Proceedings of the 11th ISCRAM Conference* (pp. 354–358).

Atefeh, F., & Khreich, W. (2015). A survey of techniques for event detection in twitter. *Computational Intelligence*, 31, 132–164.

Cameron, M. A., Power, R., Robinson, B., & Yin, J. (2012). Emergency situation awareness from twitter for crisis management. In *Proceedings of the 21st International Conference on World Wide Web* (pp. 695–698). ACM.

Caragea, C., Mcneese, N., Jaiswal, A., Traylor, G., woo Kim, H., Mitra, P., Wu, D., Tapia, A. H., Giles, L., Jansen, B. J., & Yen, J. (2011). Classifying text messages for the Haiti earthquake. In *Proceedings of the 8th International Conference on Information Systems for Crisis Response and Management (ISCRAM2011).*

Caragea, C., Silvescu, A., & Tapia, A. H. (2016). Identifying informative messages in disaster events using convolutional neural networks. In *International Conference on Information Systems for Crisis Response and Management* (pp. 137–147).

Chaudhuri, N., & Bose, I. (2019). Application of image analytics for disaster response in smart cities. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, https://doi.org/10.24251/hicss.2019.367

Daly, S., & Thom, J. A. (2016). Mining and classifying image posts on social media to analyse fires. In *Proceedings of the 13th ISCRAM Conference*. (pp. 1–14)

Dubey, R. (2019). Developing an integration framework for crowdsourcing and internet of things with applications for disaster response. In *Social Entrepreneurship: Concepts, Methodologies, Tools, and Applications* (pp. 274–283). IGI Global.

Dubey, R., Ali, S. S., Aital, P., Venkatesh, V. et al. (2014). Mechanics of humanitarian supply chain agility and resilience and its empirical validation. *International Journal of Services and Operations Management*, *17*, 367–384.

Dubey, R., Altay, N., & Blome, C. (2017). Swift trust and commitment: The missing links for humanitarian supply chain coordination? *Annals of Operations Research*, https://doi.org/10.1007/s10479-017-2676-z.

Dubey, R., Gunasekaran, A., Childe, S. J., Roubaud, D., Wamba, S. F., Giannakis, M., & Foropon, C. (2019). Big data analytics and organizational culture as complements to swift trust and collaborative performance in the humanitarian supply chain. *International Journal of Production Economics, 210,* 120–136. http://www.sciencedirect.com/science/article/pii/S0925527319300313. doi: 10.1016/j.ijpe.2019.01.023.

Dwivedi, Y. K., Shareef, M. A., Mukerji, B., Rana, N. P., & Kapoor, K. K. (2018). Involvement in emergency supply chain for disaster management: a cognitive dissonance perspective. *International Journal of Production Research, 56,* 6758–6773.

Graf, D., Retschitzegger, W., Schwinger, W., Pröll, B., & Kapsammer, E. (2018). Cross-domain informativeness classification for disaster situations. In *Proceedings of the 10th International Conference on Management of Digital EcoSystems* (pp. 183–190). ACM.

Guha-Sapir, D., Vos, F., Below, R., & Ponserre, S. (2012). *Annual disaster statistical review 2011: the number sandtrends*. Technical Report Centre for Research on the Epidemiology of Disasters (CRED).

Gupta, A., Lamba, H., Kumaraguru, P., & Joshi, A. (2013, May). Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 729–736). ACM.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9, 1735–1780.

Huang, Q., & Xiao, Y. (2015). Geographic situational awareness: mining tweets for disaster preparedness, emergency response, impact, and recovery. *ISPRS International Journal of Geo-Information, 4*, 1549–1568.

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR), 47*, 67.

Imran, M., Castillo, C., Lucas, J., Meier, P., & Vieweg, S. (2014). Aidr: Artificial intelligence for disaster response. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 159–162). ACM.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., & Meier, P. (2013a). Extracting information nuggets from disaster-related messages in social media. In *Proceedings of the 10th ISCRAM Conference,* (pp. 791–801)

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., & Meier, P. (2013b). Practical extraction of disaster-relevant information from social media. In *Proceedings of the 22nd International Conference on World Wide Web* (pp. 1021–1024). ACM.

Jabbour, C. J. C., Sobreiro, V. A., de Sousa Jabbour, A. B. L., de Souza Campos, L. M., Mariano, E. B., & Renwick, D. W. S. (2017). An analysis of the literature on humanitarian logistics and supply chain management: paving the way for future studies. *Annals of Operations Research*, 1–19. https://doi.org/10.1007/s10479-017-2536-x

Jin, S., Jeong, S., Kim, J., & Kim, K. (2015). A logistics model for the transport of disaster victims with various injuries and survival probabilities. *Annals of Operations Research, 230*, 17–33.

John, L., Gurumurthy, A., Soni, G., & Jain, V. (2018). Modelling the inter-relationship between factors affecting coordination in a humanitarian supply chain: a case of chennai flood relief. *Annals of Operations Research*, https://doi.org/10.1007/s10479-018-2963-3. doi:10.1007/s10479-018-2963-3.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kumar, A., & Singh, J. P. (2019). Location reference identification from tweets during emergencies: A deep learning approach. *International Journal of Disaster Risk Reduction*, 33, 365–375.

Kumar, A., & Singh, J. P., Saumya, S. (2019). A comparative analysis of machine learning techniques for disaster related tweet classification. *IEEE Region 10 Humanitarian Technology Conference*, (pp. 222–227).

Kumar, A., Singh, J. P., & Rana, N. P. (2017). Authenticity of geo-location and place name in tweets. In *Proceedings of the 23rd Americas conference on information systems (AMCIS)*, (pp. 1–9)

Lagerstrom, R., Arzhaeva, Y., Szul, P., Obst, O., Power, R., Robinson, B., & Bednarz, T. (2016). Image classification to support emergency situation awareness. *Frontiers in Robotics and AI, 3*, 54.

Li, H., Guevara, N., Herndon, N., Caragea, D., Neppalli, K., Caragea, C., Squicciarini, A. C., & Tapia, A. H. (2015). Twitter mining for disaster response: A domain adaptation approach. In *Proceedings of the 12th ISCRAM Conference*.

Mouzannar, H., Rizk, Y., & Awad, M. (2018). Damage identification in social media posts using multimodal deep learning, In *ISCRAM*.

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (pp. 807–814).

Nguyen, D. T., Al Mannai, K. A., Joty, S., Sajjad, H., Imran, M., & Mitra, P. (2017). Robust classification of crisis-related data on social networks using convolutional neural networks. In *11th International Conference on Web and Social Media, ICWSM 2017* (pp. 632–635). AAAI Press.

Nguyen, D. T., Alam, F., Ofli, F., & Imran, M. (2017b). Automatic image filtering on social networks using deep learning and perceptual hashing during crises. *arXiv preprint arXiv:1704.02602.*

Nguyen, D. T., Joty, S., Imran, M., Sajjad, H., & Mitra, P. (2016). Applications of online deep learning for crisis response using social media information. *arXiv preprint arXiv:1610.01030.*

Nguyen, D. T., Ofli, F., Imran, M., & Mitra, P. (2017c). Damage assessment from social media imagery data during disasters. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017* (pp. 569–576). ACM.

Olteanu, A., Castillo, C., Diaz, F., & Vieweg, S. (2014). Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *Eighth International AAAI Conference on Weblogs and Social Media*.

Papadopoulos, T., Gunasekaran, A., Dubey, R., Altay, N., Childe, S. J., & Fosso-Wamba, S. (2017). The role of big data in explaining disaster resilience in supply chains for sustainability. *Journal of Cleaner Production, 142*, 1108–1118.

Paul, J. A., & Hariharan, G. (2012). Location-allocation planning of stockpiles for effective disaster mitigation. *Annals of operations research, 196*, 469–490.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Rizk, Y., Jomaa, H. S., Awad, M., & Castillo, C. (2019). A computationally efficient multi-modal classification approach of disaster-related Twitter images. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing* (pp. 2050–2059). ACM.

Rudra, K., Banerjee, S., Ganguly, N., Goyal, P., Imran, M., & Mitra, P. (2016). Summarizing situational tweets in crisis scenario. In *Proceedings of the 27th ACM Conference on Hypertext and Social Media* (pp. 137–147). ACM.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2013). Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Transactions on Knowledge and Data Engineering, 25*, 919–931.

Shareef, M. A., Dwivedi, Y. K., Mahmud, R., Wright, A., Rahman, M. M., Kizgin, H., & Rana, N. P. (2018). Disaster management in Bangladesh: developing an effective emergency supply chain network. *Annals of Operations Research*, https://doi.org/10.1007/s10479-018-3081-y. doi: 10.1007/s10479-018-3081-y.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Singh, J. P., Dwivedi, Y. K., Rana, N. P., Kumar, A., & Kapoor, K. K. (2017). Event classification and location prediction from tweets during disasters. *Annals of Operations Research*, https://doi.org/10.1007/s10479-017-2522-3. doi: 10.1007/s10479-017-2522-3.

Sinha, A., Kumar, P., Rana, N. P., Islam, R., & Dwivedi, Y. K. (2017). Impact of internet of things (IoT) in disaster management: a task-technology fit perspective. *Annals of Operations Research*, https://doi.org/10.1007/s10479-017-2658-1, doi: 10.1007/s10479-017-2658-1.

Sundermeyer, M., Schlüter, R., & Ney, H. (2012). LSTM neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association.*

Verma, S., Vieweg, S., Corvey, W. J., Palen, L., Martin, J. H., Palmer, M., Schram, A., & Anderson, K. M. (2011). Natural language processing to the rescue? Extracting" situational awareness" tweets during mass emergency. In *Fifth International AAAI Conference on Weblogs and Social Media.*

Yu, M., Huang, Q., Qin, H., Scheele, C., & Yang, C. (2019). Deep learning for real-time social media text classification for situation awareness using hurricanes sandy, Harvey, and Irma as case studies. *International Journal of Digital Earth*, 0, 1–18. DoI: 10.1080/17538947.2019.1574316.

Zheng, X., Han, J., & Sun, A. (2018). A survey of location prediction on twitter. *IEEE Transactions on Knowledge and Data Engineering, 30*, 1652–1671.